Data Mining

Homework 6

Due: First appello (July 4): 3/7/2013, 23:59
All other appelli: 25/7/2013, 23:59

You must hand in the homeworks electronically and before the due date and time. Check the web page for instructions about collaboration, about being late, and about handing in the homework.

Problem 1. Assume that we have a memory of n bits, a set of m elements and we decide to use a Bloom filter for detecting set membership using k hash functions. Write what is the probability of a false positive, that is, what is the probability that an element that does not belong to the set is declared that it belongs.

An alternative design is similar to the one that we presented when discussing memory optimizations for the apriori algorithm used for finding frequent itemsets. We separate the memory into k distinct sections, each of n/k bits. As with the standard Bloom filter, we use k hash functions, and we allocate each section to a hash function with range $\{0, 1, \ldots, n/k - 1\}$. The algorithm is similar as before: For each of the m elements, we compute the k hash functions and we set the corresponding bit to 1, but now each hash function has its own section of the memory. To check if an element exists in the set, we compute the k hashes and for each hash we check whether the corresponding bit of the corresponding memory section is 1.

What is in this case the probability of a false positive?

Assume that you had to design a system and you have to choose one of the two types of Bloom filter (the standard one and the one presented here). Which one would you use and why?

Problem 2. An interesting phenomenon in social networks is that a random person's expected degree is smaller than the degree of her peers: "Your friends are more popular than you are!" Given an undirected graph $G = (V,E)$, select a random node and let X be the random variable that equals to its degree and Y to be the random variable that equals the average degree of the node's neighbors.

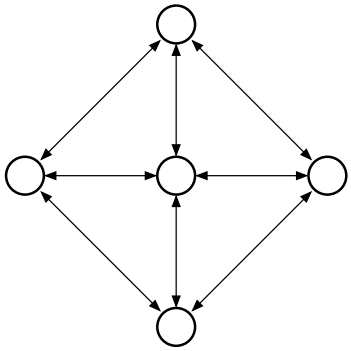
1. Prove that $\mathbf{E}[X] \leq \mathbf{E}[Y]$.
2. When do we have that $\mathbf{E}[X] = \mathbf{E}[Y]$?

Hint. Prove that for any $a,b \neq 0$ we have that $\frac{a}{b} + \frac{b}{a} \geq 2$.

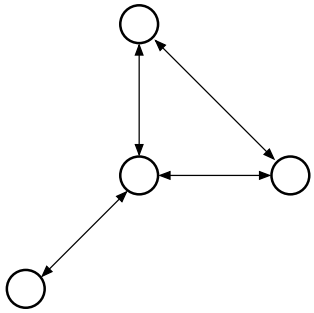
Problem 3. Here we will compute the PageRank in a special case and we will prove a rule.

1. Compute the PageRank scores of the nodes of the following two graphs, for teleporting probability equal to zero (i.e., $\beta = 1$), using the equations of the stationary distribution. Take advantage of the symmetries to reduce the number of unknown variables: are there nodes that we know a priori that they have the same PageRank score?

(A)



(B)



2. Notice that here we have a special case: All the edges are bidirectional and we have $\beta = 1$. After observing the scores of the nodes that you computed in these examples, make a conjecture about the PageRank score of a node in this special case, and prove it.

Problem 4. Recall that in the k-means problem we want to minimize the total squared ℓ_2 distance between each point and the center to which it is assigned to:

$$\sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2,$$

where C_i is the set of points that belong to the ith cluster, $\boldsymbol{\mu}_i$ the mean of the points in the ith cluster, and

$$\|\mathbf{x}\|^2 = \sum_{j=1}^{d} x_j^2,$$

if $\mathbf{x} = (x_1, x_2, \ldots, x_d)$.

In class, we said that in general the k-means problem is NP-hard. However, for $d = 1$ the problem is polynomial. Design an algorithm that solves the k-means problem in time polynomial in the number of points n and the number of clusters k, for $d = 1$.
(**Hint:** Can you solve the problem for k clusters if you assume that you can solve it for fewer than k clusters?)